A Note on the Simultaneous Computation of Thousands of Pearson's χ^2–Statistics

Holger Schwender

Collaborative Research Center 475

Department of Statistics

University of Dortmund

holger.schwender@udo.edu

Abstract

In genetic association studies, important and common goals are the identification of single nucleotide polymorphisms (SNPs) showing a distribution that differs between several groups and the detection of SNPs with a coherent pattern. In the former situation, tens of thousands of SNPs should be tested, whereas in the latter case typically several ten SNPs are considered leading to thousands of statistics that need to be computed.

A test statistic appropriate for both goals is Pearson's χ^2-statistic. However, computing this (or another) statistic for each SNP or pair of SNPs separately is very time-consuming.

In this article, we show how simple matrix computation can be employed to calculate the χ^2-statistic for all SNPs simultaneously.

1 Introduction

While association studies typically comprise the genotypes of several ten single nucleotide polymorphisms (SNPs), quite recently developed microarrays allow to measure the genotypes of tens or even hundreds of thousands of SNPs simultaneously. In the former situation, two tasks are the identification of SNPs showing a distribution that differs substantially between several groups (e.g., non-cancer vs. cancer) and the detection of groups of SNPs with a coherent pattern. Considering SNP microarray, a first goal is to reduce the number of SNPs to a better manageable size.

Since the latter goal is similar to the first task in the former situation, both problems can be solved in the same way: For each SNP, a statistic appropriate for testing if its distribution differs between several groups is computed. The higher this score, the more likely it is that the corresponding SNP differs substantially between the classes (for topics such as adjusting for multiplicity that have to be considered in this situation, see, e.g., Dudoit et al., 2003).

As SNPs are categorical variables exhibiting three realizations (homozygous reference, heterozygous, and homozygous variant), an appropriate test statistic is Pearson's χ^2-statistic

$$\chi^2 = \sum_{g=1}^{r} \sum_{k=1}^{c} \frac{(n_{gk} - \tilde{n}_{gk})^2}{\tilde{n}_{gk}}, \tag{1.1}$$

where n_{gk} and $\tilde{n}_{gk}$ are the observed and expected numbers of observations, respectively, shown in the g^{th} row and k^{th} column of the corresponding contingency table, $g = 1, \ldots, r$, $k = 1, \ldots, c$. Here, $c = 3$, and r is the number of groups such that n_{gk} specifies how many of the $n = \sum_{g,k} n_{gk} = \sum_{g,k} \tilde{n}_{gk}$ observations in the g^{th} class showing the k^{th} genotype.

A solution to the other problem, i.e. the detection of groups composed of related SNPs, is, again, to compute Pearson's χ^2-statistic (1.1) with $r = c = 3$ to test if two SNPs are independent. Afterwards, e.g., Pearson's corrected contingency coefficient

$$P_C = \sqrt{\frac{d}{d-1} \cdot \frac{\chi^2}{\chi^2 + n}}$$

with $d = \min\{r, c\} \overset{\text{here}}{=} 3$ is calculated as a measure of similarity for these two SNPs.

Since considering each pair of m SNPs means that $m(m-1)/2$ similarities/distances have to be determined, Pearson's χ^2-statistic has to be calculated several hundred to a few thousand times even if $m < 100$. The same applies to the analysis of microarrays in which tens of thousands of SNPs are tested for different group distributions. In these cases, it can therefore be time-consuming to compute each of the χ^2-statistics separately.

In this article, we show how simple matrix calculation can be employed to consider all SNPs or pairs of SNPs simultaneously. Another advantage of this approach is that it provides a matrix composed of all m or $m(m-1)/2$ contingency tables, respectively. This matrix thus enables the determination of other similarity measures based on contingency tables such as simple or flexible matching coefficients (Müller et al., 2005).

This article is organized as follows: In Section 2, the approaches for the simultaneous computation of thousands of Pearson's χ^2-statistics in the two situations are described, whereas in Section 3 we discuss practical issues as the actual implementation of these algorithms and the handling of both missing values and variables with differing numbers of levels. This section also contains a short description of how the approach for detecting groups of SNPs with coherent patterns can be used to compute similarity measures that are not based on Pearson's χ^2-statistic. Finally, the matrix based algorithms are compared with the corresponding individual computations in Section 4.

2 Simultaneous Computation of Pearson's χ^2-Statistic

Let $\mathbf{X}$ be an $m \times n$ matrix in which each column corresponds to one of the n observations and each row to one of the m variables, and $\mathbf{y}$ be a vector of length n containing the class labels of the n observations. Assume that each of these variables exhibits c levels denoted by the integers $1, \dots, c$, and that each observation belongs to one of the r classes $1, \dots, r$. Then, Pearson's χ^2-statistics for testing a variable if its distribution differs between r groups can be computed for all variables represented in $\mathbf{X}$ simultaneously by the procedure described in Algorithm 1.

Algorithm 1 (Rowwise Pearson's χ^2-Statistic)

Let $\mathbf{X}$ be an $m \times n$ matrix consisting of the values $1, \dots, c$, and $\mathbf{y}$ be a vector of length n containing the class labels $1, \dots, r$ of the n observations represented by the columns of $\mathbf{X}$.

1. Let $\mathbf{X}^{(k)}$ denote an $m \times n$ matrix with elements

$$x_{ij}^{(k)} = \begin{cases} 1, & \text{if } x_{ij} = k \\ 0 & \text{otherwise} \end{cases},$$

$k = 1, \dots, c$, and $\mathbf{L}$ be an $n \times r$ matrix with elements $\ell_{jg} = I\big(y_j = g\big)$.

2. For $k = 1, \dots c$, set $\mathbf{L}^{(k)} = \mathbf{X}^{(k)}\mathbf{L}$ and

$$\tilde{\mathbf{L}}^{(k)} = \frac{1}{n}\mathbf{X}^{(k)}\mathbf{1}_n\mathbf{L}'\mathbf{1}_n, \tag{2.1}$$

where $\mathbf{1}_n$ is a vector consisting of n ones, and compute

$$\mathbf{S}^{(k)} = \frac{\mathbf{L}^{(k)} * \mathbf{L}^{(k)}}{\tilde{\mathbf{L}}^{(k)}} \tag{2.2}$$

with $s_{ig}^{(k)} = \ell_{ig}^{(k)} \cdot \ell_{ig}^{(k)} \,/\, \tilde{l}_{ig}^{(k)}$, $i = 1, \ldots, m$, $g = 1, \ldots, r$.

3. Let $\mathbf{S}$ be an $n \times c$ matrix in which the k^{th} column consists of the vector $\mathbf{s}_k = \mathbf{S}^{(k)}\mathbf{1}_r$, $k = 1, \ldots, c$. The vector comprising Pearson's χ^2-statistics for testing each row of $\mathbf{X}$ if the distribution of the corresponding variable differs between the groups specified by $\mathbf{y}$ is given by

$$\mathbf{r}_{\mathbf{X}} = \mathbf{S}\mathbf{1}_c - n. \tag{2.3}$$

For the computations in Algorithm 1, note that (1.1) can also be expressed as

$$\chi^2 = \sum_{g=1}^{r} \sum_{k=1}^{c} \frac{n_{gk}^2}{\tilde{n}_{gk}} - n,$$

and that the $\left(i^{\text{th}}, g^{\text{th}}\right)$ element of $\mathbf{L}^{(k)}$ or $\tilde{\mathbf{L}}^{(k)}$ comprises the observed or expected number of observations, respectively, being a member of group g and showing the k^{th} level at the i^{th} variable.

Using similar ideas, Algorithm 2 describes how Pearson's χ^2-statistic for testing if two variables are independent can be determined for all of the $m(m-1)/2$ pairs of m variables simultaneously. Note that Algorithm 2 assumes that all variables exhibit the same number of levels such that $r = c$. For an extension to $r \leq c$, see Section 3.3.

Algorithm 2 (Pairwise Pearson's χ^2-Test for Independence)

Let $\mathbf{X}$ be an $m \times n$ matrix consisting of the values $1, \ldots, c$.

1. Let $\mathbf{X}^{(k)}$ denote an $m \times n$ matrix, $k = 1, \ldots, c$, with elements

$$x_{ij}^{(k)} = \begin{cases} 1, & \text{if } x_{ij} = k \\ \\ 0 & \text{otherwise} \end{cases}.$$

2. For $g, k = 1, \ldots, c$, compute $\mathbf{N}^{(gk)} = \mathbf{X}^{(g)}\mathbf{X}^{(k)\prime}$ and

$$\tilde{\mathbf{N}}^{(gk)} = \frac{1}{n}\mathbf{X}^{(g)}\mathbf{1}_n\mathbf{1}_n'\mathbf{X}^{(k)\prime}. \tag{2.4}$$

3. Pearson's χ^2-statistic for testing if the i^{th} and the h^{th} variable, $i, h = 1, \ldots, m$, represented in the i^{th} and h^{th} row of $\mathbf{X}$, respectively, are independent is given by the $\left(i^{\text{th}}, h^{\text{th}}\right)$ element of

$$\mathbf{R}_{\mathbf{X}} = \sum_{g=1}^{c}\sum_{k=1}^{c} \frac{\mathbf{N}^{(gk)} * \mathbf{N}^{(gk)}}{\tilde{\mathbf{N}}^{(gk)}} - n. \tag{2.5}$$

3 Practical Issues

3.1 Details on the Actual Implementation

Algorithms 1 and 2 have been implemented in the statistical software environment `R` (Ihaka and Gentleman, 1996). The following notes give details on their actual implementations:

- Instead of computing the rowwise or columnwise sums of a matrix by multiplying it with a vector consisting of an appropriate number of ones, the even faster `R` function `rowSums` or `colSums` are employed for these calculations.

- In (2.2) and (2.5), the elementwise squaring of $\mathbf{L}^{(k)}$ or $\mathbf{N}^{(gk)}$ is represented by “*”, since in `R` it is faster to elementwise square a matrix $\mathbf{Z}$ by $\mathbf{Z} * \mathbf{Z}$ than by $\mathbf{Z}^2$. Thus, the actual reason for this is not to avoid notations such as $\mathbf{N}^{(gk)2}$.

- In the second step of Algorithm 2, not all c^2 matrices $\mathbf{N}^{(gk)}$ are determined. We only consider $\mathbf{N}^{(gk)}$ for $g = 1, \ldots, c$, and $k = g, \ldots, c$, as the lower (upper) triangle of $\mathbf{N}^{(gk)}$ contains the same values as the upper (lower) triangle of $\mathbf{N}^{(kg)}$. The same applies to $\tilde{\mathbf{N}}^{(gk)}$.

- Finally, the pairwise χ^2-statistics are actually not computed as shown in (2.5). Instead, the upper triangle of $\mathbf{N}^{(gk)}$ is stored in the $\big((g-1)c+k\big)^{\text{th}}$ column of the $m(m-1)/2 \times c^2$ matrix $\mathbf{M}$, and the lower triangle in the $\big((k-1)c+g\big)^{\text{th}}$ column. Hence, this matrix $\mathbf{M}$ contains all contingency tables corresponding to any of the $m(m-1)/2$ pairs of m variables. Therefore, $\mathbf{M}$ can also be employed to compute other similarity measures based on contingency tables. In the same way, all expected cell entries are retained in a matrix $\tilde{\mathbf{M}}$.

  The vector $\mathbf{r}_{\mathbf{X}}$ comprising the $m(m-1)/2$ χ^2-statistics can thus be determined by
$$\mathbf{r}_{\mathbf{X}} = \left(\frac{\mathbf{M} * \mathbf{M}}{\tilde{\mathbf{M}}} - n \right) \mathbf{1}_{c^2}.$$
  By default, $\mathbf{r}_{\mathbf{X}}$ is then stored in the lower triangle of an $m \times m$ matrix such that the lower triangle of this matrix is identical to the lower triangle of $\mathbf{R}_{\mathbf{X}}$ in (2.5).

3.2 Missing Values

Even though both Algorithm 1 and 2 still work if there are missing values, the resulting χ^2-statistics will not be correct, since the expected numbers of observations are divided by n, i.e. the total number of observations, and not by the actual number of observations showing no missing value at a particular variable.

For solving this problem, let $\mathbf{X}^{\mathrm{A}}$ be an $m \times n$ matrix with

$$x_{ij}^{\mathrm{A}} = \begin{cases} 1, & x_{ij} \text{ is not missing} \\ 0, & x_{ij} \text{ is a missing value} \end{cases},$$

and replace n in (2.3) by the vector

$$\mathbf{n} = \begin{bmatrix} n_1 \\ \vdots \\ n_m \end{bmatrix} = \mathbf{X}^{\mathrm{A}} \mathbf{1}_n.$$

In (2.1), $\frac{1}{n}$ is substituted by the $m \times r$ matrix

$$\mathbf{N}^{\mathrm{den}} = \begin{bmatrix} \frac{1}{n_1} & \cdots & \frac{1}{n_1} \\ \vdots & \ddots & \vdots \\ \frac{1}{n_m} & \cdots & \frac{1}{n_m} \end{bmatrix},$$

whereas in (2.4) and in (2.5) n is replaced by the $m \times m$ matrix

$$\mathbf{N} = \mathbf{X}^{\mathrm{A}} \left(\mathbf{X}^{\mathrm{A}}\right)'.$$

Since (2.4) just takes individual and not pairwise missing values, i.e. missing values appearing in either of the two considered variables, into account, it is additionally necessary to replace the rowwise sums of $\mathbf{X}^{(k)}$ by $\mathbf{Z}^{(k)} = \mathbf{X}^{(k)} \left(\mathbf{X}^{A}\right)'$ such that (2.4) becomes

$$\tilde{\mathbf{N}}^{(gk)} = \frac{\mathbf{Z}^{(g)} * \left(\mathbf{Z}^{(k)}\right)'}{\mathbf{N}}.$$

3.3 Different Numbers of Levels

If the i^{th} variable exhibits $r < c$ levels, where c is the maximum number of levels that one of the m variables can take, then the i^{th} row of $\mathbf{X}^{(r+1)}, \dots, \mathbf{X}^{(c)}$ will only consist of zeros. Hence, the i^{th} row of $\mathbf{L}^{(k)}$ and $\tilde{\mathbf{L}}^{(k)}$ in Algorithm 1 and the corresponding rows and columns of $\mathbf{N}^{(gk)}$ and $\tilde{\mathbf{N}}^{(gk)}$, $k = r+1, \dots, c$, or $g = r+1, \dots, c$, in Algorithm 2 will be composed of zeros. Since this leads to dividing zero by zero in (2.2) and (2.5), no χ^2-statistic for the i^{th} variable will be available.

A solution to this problem is to set $\tilde{\ell}_{ig}^{(k)} = \max\Big\{1,\ \tilde{\ell}_{ig}^{(k)}\Big\}$ in $\tilde{\mathbf{L}}^{(k)}$, and $\tilde{n}_{ih}^{(k)} = \max\Big\{1,\ \tilde{n}_{ih}^{(k)}\Big\}$ in $\tilde{\mathbf{N}}^{(k)}$.

3.4 Computation of Similarity Measures

As mentioned in Section 3.1, in the actual realization of Algorithm 2, an $m(m-1)/2 \times c^2$ matrix $\mathbf{M}$ containing all $m(m-1)/2$ contingency tables of the pairwise comparisons is constructed that enables the computation of similarity measures such as the simple matching coefficient

$$S_M = \frac{1}{n}\sum_{k=1}^{c} n_{kk}.$$

Since the $\big((k-1)c+h\big)^{\text{th}}$ column of $\mathbf{M}$ comprises the entry n_{kh} of each of the contingency tables, the vector $\mathbf{r}_S$ consisting of S_M for any pairwise comparison of two variables is given by

$$\mathbf{r}_S = \frac{\mathbf{Md}}{\mathbf{M1}_{c^2}},$$

where $\mathbf{d}$ is a vector of length c^2 with elements

$$\mathbf{d}_h = \begin{cases} 1, & \text{if } h \in \big\{a:\ a = (k-1)c+k,\ k = 1, \dots, c\big\} \\ 0 & \text{otherwise} \end{cases}.$$

4 Discussion

In this paper, we have presented approaches based on matrix algebra for the simultaneous computations of thousands of Pearson's χ^2-statistics.

Table 4.1 shows that the computation time of Algorithm 1 depends only slightly on the number of classes. However, the number of levels that a variable can take has a high influence on the computation time time. This is not very surprising, as the larger c, the more matrices $\mathbf{X}^{(k)}$ have to be constructed and evaluated in the determination of the χ^2-statistics.

TABLE 4.1. Computation times of Algorithm 1 for different numbers m of variables, numbers c of levels a variable can take, and numbers r of classes to which the $n = 200$ observations belong.

	$r = 2$,	$r = 2$,	$r = 2$,	$r = 3$,	$r = 6$,
m	$c = 3$	$c = 5$	$c = 10$	$c = 3$	$c = 3$
100	< 0.01	0.01	0.01	< 0.01	0.01
1,000	0.05	0.07	0.15	0.05	0.05
10,000	0.63	1.03	2.04	0.64	0.62
100,000	6.16	9.98	61.82	6.18	6.46

As Table 4.2 and 4.3 reveal, using Algorithms 1 and 2 lead to a substantial decreases of the computation time in comparison to one-by-one determinations of the χ^2-statistics. Not very surprisingly, the more χ^2-statistics, the higher is the factor by which the computation is accelerated. But even if the number of variables is small, the algorithms will be about 15 times faster.

Algorithm 1 is used in version 1.10.0 and later of the R package `siggenes` available at `http://www.bioconductor.org`, the web page of the Bioconductor project (Gentleman et al., 2004), such that the computation time of both

TABLE 4.2. Computation times of both Algorithm 1 and the individual calculation of Pearson's χ^2-statistics for different numbers m of variables and numbers n of observations. Each variable can take $c = 3$ levels, and each observation belongs to one of $r = 2$ classes.

	Algorithm 1		Individual	
m	$n = 200$	$n = 1,000$	$n = 200$	$n = 1,000$
50	< 0.01	0.01	0.13	0.16
100	< 0.01	0.02	0.26	0.32
1,000	0.05	0.40	2.64	3.35
10,000	0.63	2.39	26.74	34.42
100,000	6.16	–	274.96	–

TABLE 4.3. Computation times of both Algorithm 2 and the individual calculation of Pearson's χ^2-statistic for testing each pair of m variables if they are independent, where each variable exhibits $c = 3$ levels, and the number of observations is $n = 1,000$.

m	Algorithm 2	Individual
10	0.01	0.15
50	0.07	4.25
100	0.33	17.32
200	1.22	70.06
500	7.79	474.22

SAM (Significance Analysis of Microarrays; Tusher et al., 2001) and EBAM (Empirical Bayes Analysis of Microarrays; Efron et al., 2001) applied to categorical data (Schwender, 2005, 2007) is reduced. Both Algorithm 1 and 2 are implemented in the R function `rowChisqStats` contained in the package `scrime` that will be available soon at http://www.bioconductor.org and http://cran.r-project.org.

Acknowledgements

The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of Complexity in Multivariate Data Structures") is gratefully acknowledged.

References

DUDOIT, S., SHAFFER, J. P., and BOLDRICK, J.C. (2003). Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science*, **18**(1), 71–103.

EFRON, B., TIBSHIRANI, R., STOREY, J.D., and Tusher, V. (2001). Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*, **96**, 1151–1160.

GENTLEMAN, R.C., CAREY, V.J., BATES, D.M., BOLSTAD, B., DETTLING, M., DUDOIT, S., ELLIS, B., GAUTIER, L., GE, Y., GENTRY, J., HORNIK, K., HOTHORN, T., HUBER, W., IACUS, S., IRIZARRY, R., LEISCH, F., LI, C., MAECHLER, M., ROSSINI, A.J., SAWITZKI, G., SMITH, C., SMYTH, G., TIERNEY, L., YANG, J.Y.H., and ZHANG, J. (2004). Bioconductor: Open Software Development for Computational Biology and Bioinformatics. *Genome Biology*, **5**, R80.

Ihaka, R., and Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314.

Müller, T., Selinski, S., and Ickstadt, K. (2005). Cluster Analysis: A Comparison of Different Similarity Measures for SNP Data. *Technical Report*, SFB 475, Department of Statistics, University of Dortmund.

Schwender, H. (2005). Modifying Microarray Analysis Methods for Categorical Data – SAM and PAM for SNPs. In Weihs, C., and Gaul, W. (eds.), *Classification – The Ubiquitous Challenge*. Springer, Heidelberg, 370–377.

Schwender, H. (2007). Empirical Bayes Analysis of Single Nucleotide Polymorphisms. *In Preparation.*

Tusher, V., Tibshirani, R., and Chu, G. (2001). Significance Analysis of Microarrays Applied to the Ionizing Radiation Response. *Proceedings of the National Academy of Sciences*, **98**, 5116–5124.